CHAPTER V
CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

Chapter Five presents the research summary and the summary of the findings in the first part. The conclusions and the implications for language testing are presented in the second part. Subsequently, recommendations for future research are also made.

5.2 Research summary and summary of the findings

5.2.1 Research summary

This study concerns the investigation of the effects of the two independent variables: test authenticity and test delivery mediums on test takers' reading proficiency. Furthermore, the attitudes of the test takers towards test authenticity and test delivery mediums were explored.

This study emphasizes the reading comprehension tests tailored for the first year students at King Mongkut's Institute of Technology North Bangkok. For this purpose, the test authenticity in this study was defined by the target language use (TLU) domain. The TLU domain was specified by the characteristics obtained from a survey conducted with English language instructors, currently-enrolled students, instructors of the other faculties at King Mongkut's Institute of Technology North Bangkok and employers of graduates from King Mongkut's Institute of Technology North Bangkok. Therefore, the authentic tests developed in this study were considered more relevant to their real-life language use.

Another investigated variable was test delivery mediums consisting of conventional paper-and-pencil test administration and computer-based administration. The type of computer-based tests created in this study was Computer-Adaptive Test (CAT). The model used was a fixed-branch computer-adaptive test. All the items in the test are placed in a branching tree in advance. A different item is chosen to be presented next, a more difficult item for a correct response and an easier item for an incorrect response according to the fixed-branch established. Moreover, the CATs in this study were content-based CATs. Therefore, an incorrect response to a previous item would lead to an easier item measuring the same dimension. On the other hand, a correct response to a previous item would introduce a more difficult item measuring a different dimension.

Consequently, the research instruments used in this study were:

1. Paper-based tests

The Authentic English Reading Comprehension Conventional Paper-and-Pencil Test (ACON) is a paper-and-pencil test. The item types used in this test are short answer, gap-filling and information-transfer.

The Inauthentic English Reading Comprehension Conventional Paper-and-Pencil Test (ICON) is different form ACON in terms of the item type used. All the items in this test are written in the form of multiple-choice format.

2. Computer-based tests

The Authentic English Reading Comprehension Computer-Adaptive Test (ACOM) is a computer-based test. The items are from ACON. The item types used in this test are short answer, gap-filling and information-transfer. The test items are tailored according to the test-takers' ability by using constant six-step pyramidal model of CAT.

The Inauthentic English Reading Comprehension Computer-Adaptive Test (ICOM) is different from ACOM in terms of the item type used. All the items in this test are written in multiple-choice format.

3. The interview schedule was conducted by means of retrospective method.

There are two research questions in this study. They are:

1. Can test authenticity and test delivery mediums have any effect on test takers' English reading proficiency and what are their effect sizes?

2. What are the test takers' attitudes towards test authenticity and test delivery mediums?

This study was conducted with 300 first year students at King Mongkut's Institute of Technology North Bangkok in the second semester of the academic year 2006. The students studied in the Faculties of Engineering, Applied Science, Technical Education, Industrial Technology and Management and Agro-Industry. Typically, most were male, aged between 18 – 22.

5.2.2 Summary of the findings

Two-way Analysis of Variance (2*2 ANOVA) was used to identify the main effects and the interaction effects of the two variables: test delivery mediums and test authenticity. Furthermore, Partial Eta squared reported by SPSS program was used to

measure the effect sizes of these variables. The results of this analysis can justify the research hypotheses as follows:

Hypothesis 1: the mean score obtained from the authentic reading comprehension test is significantly different from that obtained from the inauthentic reading comprehension test at the significant level of .05 (Statistical hypothesis is $\overline{X}$ Authentic ≠ $\overline{X}$ Inauthentic).

The 2*2 ANOVA reveals that there was a significant difference between the mean score obtained from the test considered authentic (ACOM and ICOM) and those obtained from the test considered inauthentic (ICOM and ICON) at the 0.05 level ($p < .05$, $F = 49.164$).

Partial Eta Squared value (0.142) shows the medium effect size of the test authenticity on test takers' reading ability score.

Hypothesis 2: the mean score obtained from the computer-based reading comprehension test is significantly different from that obtained from the paper-based reading comprehension test at the significant level of .05 (Statistical hypothesis is $\overline{X}$ CBT ≠ $\overline{X}$ PBT).

The results of the 2*2 ANOVA indicates that there was no significant difference between the mean score obtained from the test delivered by means of computer (ACOM and ICOM) and that obtained from the test delivered by means of paper (ACON and ICON) at the 0.05 level ($p > .05$, $F = 2.556$).

The effect size value of test delivery mediums was 0.009. This means that the effect size of the mediums on test takers' reading ability scores is small.

Hypothesis 3: there are significant interaction effects between test authenticity and test delivery mediums on students' reading proficiency at the significant level of .05. (Statistical hypotheses are:

$\overline{X}$ Authentic CBT ≠ $\overline{X}$ Inauthentic CBT

$\overline{X}$ Authentic PBT ≠ $\overline{X}$ Inauthentic PBT

$\overline{X}$ Authentic CBT ≠ $\overline{X}$ Authentic PBT

$\overline{X}$ Inauthentic CBT ≠ $\overline{X}$ Inauthentic PBT

$\overline{X}$ Authentic CBT ≠ $\overline{X}$ Inauthentic PBT

$\overline{X}$ Inauthentic CBT ≠ $\overline{X}$ Authentic PBT).

The interaction effect between Mediums and Authenticity value obtained from the 2*2 ANOVA was significant at the level of 0.05 ($p \leq .05$, $F = 14.474$).

Furthermore, a descriptive technique was conducted to examine the test takers' attitudes. The highest frequency of the attitudes towards test authenticity and test delivery mediums was reported. The chi square (χ^2) test was applied to find out whether there was any significant difference in the proportions of attitudes towards test authenticity and test delivery mediums in each group.

The findings of the interviews can be categorized as follows:

1. Similar attitudes among the four groups:

1.1 Similar positive opinions

The four groups (ACOM, ACON, ICOM and ICON) reported positive opinions on the following issues:

1) the opportunity to demonstrate strengths and weaknesses in reading ability

2) the perception of the test fairness: the appropriateness of the question

3) the perception of nervousness while taking the tests

4) the accuracy of the test to elicit their true ability in reading

5) the accuracy of the test to elicit their true ability in reading: the relevance of the test tasks to their real life reading activities

6) test administration of the test they took

7) test administration: compared to other versions of the tests

8) facilities of the test delivery mediums (PBT and CBT)

9) familiarity with the test delivery mediums (PBT and CBT) and the test authenticity (short answer and multiple choice item types)

10) perseverance, and

11) attitudes.

1.2 Similar neutral opinions

The only similar opinion reported by the four groups was on test difficulty.

1.3 Similar negative opinions

The four groups reported negative opinions on the perception of the test fairness: the effects of the method to answer the test.

2. Different attitudes reported by the four groups

The following topics were reported with different opinions among the four groups:

1) the perception of test fairness: the situations in the tests were reported differently. ACOM (f=3), ICOM (f=3), ACON (f=3) tended to have positive opinions but ICON reported (f=5) negative opinions.

2) time length: the three groups, ACOM (f=5), ICOM (f=5) and ICON (f=5) totally had positive opinions on this topic but the only one test taker in ACON reported a positive opinion whereas the rest (f=4) reported negative opinions.

3) test contents: 100 per cent of ACOM (f=5) , ICOM (f=5) and ACON (f=5) reported positive opinions while only two of the five in ICON (f=2) reported positive and three reported neutral opinions.

4) scoring: Most of the frequencies of ACOM (f=4) , ICOM (f=5) and ICON (f=3) reported positive opinions. Conversely, 100 per cent of ACON (f=5) reported negative opinions.

5) interactiveness: Most of the frequencies of ACOM (f=3) reported neutral opinions. One hundred per cent of ACOM reported neutral opinion. Most of the frequencies of ACON (f=3) reported positive opinions and the rest (f=2) reported neutral opinions. Three frequencies of ICON reported positive opinions and two frequencies reported negative opinions.

6) authenticity: All frequencies of ACOM (f=5) reported a positive opinion on this topic. Three frequencies of ICOM reported positive opinions and two frequencies reported negative opinions. The majority of ACON (f=3) reported positive opinions and two reported neutral opinions. Lastly, only one frequency of ICON reported a positive opinion while the majority reported negative opinions.

5.3 Conclusions and implications of the study

5.3.1 Conclusions

5.3.1.1 Test authenticity

Regarding the current teaching approach, constructivism, test authenticity is considered one of the three main important features (Jonassen, 1991). From this perspective, it can be said that authenticity is the focus of language learning and teaching.

In terms of test development, test writers should construct a test which can be expected to show the influence of current ideas on what constitutes language ability and what exactly we are doing when we use language in our everyday life. Hence, McNamara (2000) mentions that in major test projects, articulating and defining the test construct may be the first stage of test development, resulting in an elaborated statement of the theoretical framework for the test.

With this respect, test authenticity can be determined by the construct of the test as Bachman and Palmer (1996) suggest that construct validity is relevant to the domain of language use to which our score interpretations generalize. The domain of generalization is the set of tasks in the Target Language Use (TLU) domain to which the test tasks correspond. In defining the TLU domain, Bachman and Palmer (1996) mention that the response format is another factor to consider in test design in order to increase test authenticity. The test design tends to move toward what is variously known as alternative, or standards-based assessment, which includes judging students' ability to perform more open-ended, holistic and real-world tasks within their normal learning environment.

Moreover, in terms of having the more authentic tests, Barton's study (1994) shows the richness of the social world within which literacy events take place. Reading will often be accompanied by talking like reading aloud a snippet from a newspaper in order to discuss a political bias or the performance of a football team. Similarly Alderson (2000) suggests that reading should be assessed within a number of situations.

The authentic reading tests in this study were therefore constructed by the students' TLU domain while the task types allowed more open-ended responses within provided situations. The tasks in these tests corresponded to the important attributes of authentic reading comprehension tests (Mueller, 2003). The reading tasks allowed the test takers to demonstrate their understanding in more open-ended ways as in real life. They had a chance to write down their own answers without forced choice. And the answers were corrected by providing more acceptable answers. These reading tasks tended to occur in everyday life before or after students graduated. During an authentic reading test, a test taker, therefore, had more meaningful purposes for reading such as reading for a job application, recreation or obtaining important data to prepare an academic report. When taking the reading tests, it was assumed that test takers applied their understanding of real world tasks and replicated them meaningfully. Finally, test takers themselves were

involved in the test design. The reading topics and the characteristics of reading tasks were obtained from a needs analysis of test takers.

Conversely, the tests considered inauthentic in this study lacked authenticity caused by the impoverished contexts within the tests, and the limited range of situations in the test.

The findings show there was a significant difference between the mean scores obtained from the tests considered authentic and those obtained from the tests considered inauthentic. The findings lend support to the results of Wagner's study (2002) which focuses on two authentic variables: the mediums used to deliver the aural input in the listening test and test item types. He explored the listening process when the aural input was delivered through the use of the video which was considered a more authentic test. Two types of test items used in the test were limited production item types, considered as a more authentic item type and multiple-choice item types, considered as a less authentic item type. The results seem to provide some evidence for the effects of the test method on the test takers' performance. The findings indicate that the limited production item type may be more suitable for testing a listener's ability to comprehend inferential information, while the multiple-choice item type may be better suited to assess a listener's ability to comprehend explicitly stated information.

It is evident that different types of test items, which signify different degrees of test authenticity (the limited production item type as a more authentic item and the multiple-choice item type as a less authentic item type) can affect test takers' listening ability. Moreover, the more authentic item types are likely to access the higher level of listening ability.

Similarly, Lynch (2003) studied authentic performance-based assessment in ESL/EFL reading instruction and found the effects of test authenticity on test takers' performance. A reading exercise was implemented as a representative instructional model and it was used to inform a valid performance-based reading comprehension test. The performance-based reading assessment tasks were considered authentic since they involved reading for authentic purposes and the selection of texts depended on the students being assessed as well as the specific domain characteristics of the context within which they were expected to perform in studying or work. He reported that the advantage of performance-based testing is that it can create and maintain a positive washback on the teaching and learning process. Moreover, a comprehensively valid interaction between the nature of the instruction preceding the evaluation and the actual

performances being assessed was found. It can be summarized that authenticity can allow more validity in performance-based reading comprehension tests.

Obviously, the results of Wagner's study (2002) and Lynch's study (2003) support the use of authenticity in language assessment. Regarding Wagner's study (2002), more authentic item types can affect test takers' listening ability and are likely to access the higher level of listening ability. In terms of reading tests, Lynch's study (2003) illustrates that the authentic reading tests can allow more valid performance-based reading comprehension tests.

It can be concluded that the mean scores obtained from the test takers who took different tests with different degrees of test authenticity were significantly different because of the characteristics of the item types. The more authentic item type used in this study (short answers) tends to assess higher levels of reading ability. Therefore, the mean scores obtained from the ACOM and ACON groups were lower than those obtained from the ICOM and ICON groups. However, the results of the authentic tests tended to be more valid because the test takers were given greater opportunities to respond to the tasks considered more relevant to their real life whereas the multiple choice items in the inauthentic tests provided the test takers with a only limited chances to respond to those items. Only one correct answer to each task was required in the inauthentic tests while there were more acceptable answers in the authentic tests.

5.3.1.2 Test delivery mediums

Reading on computer screens is becoming more common in our daily life as the amount of reading material available online is rapidly increasing. This trend is evident in the field of language assessment where computerized testing, such as computer-based tests is attracting the attention of researchers, language learners, and test users (Kaya-Carton, Carton and Dandolini, 1991).

The findings of this study show no significant difference between the mean scores obtained from the test delivered by means of computer and that delivered on paper. The findings were similar to those in many studies described below.

Shaw et al. (2001) and Thighe et al. (2001) investigated the equivalence of PB and CB forms of the Listening and Reading Modules of IELTS. They reported that correlations of 0.83 and 0.90 were found between scores on the computer-based (CB) and paper-based (PB) versions of the Listening and Reading Modules respectively, satisfying

the criterion of 0.80 and suggesting that format had a minimal effect on the awarded scores.

Jorgensen (2003) compared the performance of the test takers on paper-based and computer-based tests of the English diagnostic test. He found that there was no significant difference between the scores of both tests. The similarity of the test takers' performance in both paper-based and computer-based tests can be described in terms of the test format. Both tests consisted of multiple-choice questions that measured the examinees' ability to understand spoken English, or English reading passages, and recognized correct English grammar. Both tests also required the examinees to write an essay.

However, no findings of significant difference were found to show that the two tests were equivalent. Chalhoub-Deville and Deville (1999) point out the importance of conducting comparability studies to detect any potential test delivery medium effect when a conventional test is converted to a computerized test. This is because the presence of a mode effect on reading comprehension test performance would seriously invalidate score interpretation of computerized reading tests. Alderson (2000) suggests that language assessment researchers have to discuss the necessity of examining: (a) the degree to which computerized reading comprehension tests measure the same construct as paper-based tests and (b) the extent to which results of computerized reading tests can be generalized to other contexts.

Due to the comparability of tasks controlled at the beginning of the test construction, it was possible to establish the equivalence of computerized and conventional test forms. The content covered by the two tests was comparable. Using the fixed length CAT in this study, promising algorithms to control content coverage have been implemented.

Bugbee (1996) claimed that if a CBT is used as an alternative for a conventional form, then demonstrating a high correlation and nearly equal means and variances between the modes may suffice. Correspondingly, the findings indicate no significant difference between the mean scores obtained from the tests delivered by these two modes. It can be said that the mean scores are nearly equal. This can be interpreted as an indication that the computer-based tests in this study can be alternatives for a paper-based test.

In terms of the equivalence of item difficulty parameters obtained in the paper-based and computer-based (CAT) forms, Stone and Lunz (1994) found that the

text-only items showed a strong trend of parameter estimate equivalence, items with graphics tended to be less stable than text-only items. Further investigation of the test items suggested that the significantly different difficulty estimates obtained across modes seemed to be accounted for by the different picture quality as well as by image and character sizes used across the CAT mode. Hence, it can be concluded that the item difficulty parameters in both paper-based and computer-based tests were generally stable since all items were text-only.

Accordingly, the finding of no significant difference in the mean scores of the tests delivered by different mediums may suggest that the construct being measured by the tests administered in conventional and computerized forms is comparable.

It can be concluded from the findings that the mean scores obtained from computer-based and paper-based tests are not significantly different because the tests are comparable. A possible cause might be that both tests were developed from the same constructs, so there is stability of the test difficulty parameter and the similar values of the mean scores.

5.3.1.3 Interaction effects

Although no main effects of test delivery medium and test authenticity were found, the interaction effects were significant. This finding is supported by Alderson (2000), who comments that the effects of test methods (CBT, PBT) and item types used such as the multiple-choice technique and other more innovative types are interesting to explore so they might affect test takers' ability in various ways. The results of this study reveal that a high degree of test authenticity causes lower scores in the tests delivered by computer. Conversely, a low degree of test authenticity or the tests considered inauthentic in this study allows higher scores in the test delivered by the computer. Where tests delivered on paper are concerned, test takers tend to get higher scores when they take the tests with low degree of test authenticity (the inauthentic tests) and vice versa.

5.3.1.4 Attitudes

Apart from the effects of medium and test authenticity on test takers performance and attitudes, another concern related to the construct validity of the tests is the effects of examinees' attitudes toward new forms of language tests. An investigation of these issues is important because a test score of computer-based tests or paper-based

tests should reflect the construct of interest only. Therefore, if the test score represents language ability then a valid generalization of test scores across modes is possible.

From the findings, it can be concluded that test takers tended to have positive attitudes towards the tests they took. However, there are some issues to which they reported negative attitudes.

The four groups reported negative opinions on the perception of test fairness: the effects of different methods in answering the test. The qualitative data from the interview shows that the majority of students who took ACOM reported negative opinions. In order to respond to test questions, they were required to type a word or phrase in the provided spaces. The findings on response methods were similar to those in Russell's study (1999). In this study, reading performance of middle-school students was controlled. The results show that students with low keyboarding speed were disadvantaged by a computer-writing test relative to students with similar low levels of keyboarding skills taking a paper test. The opposite effect was detected in students with high keyboarding speed, who were better on the computer than on paper examinations.

Therefore, in terms of the ICOM (a multiple choice computer-based test), the findings indicate that test takers tended to respond favorably to the test when demands placed upon them were limited: for example, they simply clicked on the answer in multiple-choice tests. Where demands were greater, such as having to type in words or phrases, test takers tended to react much less favorably to the test, with a preference for a paper-based version of the test (ICON and ACON). These findings lend support to those obtained from Madsen (1991) and Kiratibodee (2006), who reported on test takers' anxiety as a drawback of tests delivered by the computer. Another drawback is a potential bias in computerized exams due to unfamiliarity with the new technology. Interaction with the computer may thus be a stressful experience for some. Accordingly, test takers tend to have negative attitudes towards the test requiring them to provide more open-ended answers.

The results show that the scoring method and the time length reported by ACON test takers were negative. Since they were allowed to write more open-ended answers, they reported that they were not certain about the marking scheme: their scores might depend on either the teacher or the marker and the scoring method might not be standard. Moreover, they spent more time writing down the answers. Additionally, the majority of this group reported that they had insufficient time for taking the test.

Interestingly, 100 per cent of ACOM and ICOM (computer-based tests) test takers reported positive attitudes toward the time length. The findings lend support to findings from a study by Sukamolson (1993), where the experimental findings reveal CBT superiority to paper-and-pencil tests in terms of reliability and validity, particularly when relatively few items are administered and there is a substantial reduction in time for the exam.

There are crucial findings of the interview. Since the selection of reading texts was guided by the data obtained from the survey; the three groups (ACOM, ICOM and ACON) reported positive attitudes towards them. They also tended to have positive attitudes towards interactiveness and test authenticity. These findings support Bailey's (1999) conclusion that involvement of test takers in test construction promotes their perception of tests as more interactive and authentic, thus increasing motivation and possibly enhancing preparation and performance.

5.3.2 Implications of the Study

Since the main effects of test authenticity and interaction effect hypotheses are confirmed by findings, the implications of which are set out as follows:

1. Theoretically and practically, the findings will contribute to more meaningful constructs of a reading comprehension test. Test developers have to consider the target language use in real life in the test items. Moreover, the score obtained has greater generalisability and is more meaningful.
2. Target language use or test authenticity may not be the only important aspect which affects test takers' scores. Other variables have to be explored.
3. In terms of technology used for language assessment, Bernhardt (1991) recommends technology for its capacity to trace test takers' language development thus enabling researchers to better understand how aspects of the construct evolve across different ability levels.
4. Regarding the predictive validity of the reading comprehension, since the test scores obtained from the test are related to the target language use in real life, they may predict some future behaviors of the test takers.
5. Findings relating to the effects of test authenticity and mediums of test delivery can provide information for educational practitioners to develop more effective reading comprehension tests. However, attention must be paid to some critical features of reading performance captured in computer simulations. Also careful interpretations about learners'

knowledge and abilities should be made with caution. Language test developers may ask themselves what kinds of interpretations are actually needed to describe reading abilities; and what kinds of simulation tasks will provide the required evidence.

6. Regarding modes of presentation, presenting reading texts and items using the computer can be a supplement or a replacement of an existing test. However, revisions of test specifications, test format and layout with enhanced authenticity, construct validity, and measurement accuracy must be carefully carried out.

7. The test takers' attitudes towards test authenticity and means of test delivery show that student perceptions differed on the issues of test fairness (the situations in the tests); time length; scoring method; test contents, interactiveness and test authenticity. Therefore, test developers should pay an attention to these factors when constructing reading comprehension tests.

However, in this study there were no findings on the main effects of test delivery mediums. Therefore, the findings provide educators and researchers with the following insights:

1. Target language use domains have to be carefully defined. Though the survey of TLU domain was conducted on the majority of the students, the data may be different for students from different groups. In this way, some students may report their perception of the test authenticity differently.

2. Test authenticity and mediums of test delivery may not be important variables in assessing reading comprehension. Other human factors like perception of test fairness should be investigated.

3. Test takers' attitudes towards test authenticity and mediums of test delivery may not be a crucial factor affecting reading performance. On the other hand, other variables like students' confidence in using the computer may have some effects. Such effects were indicated in Brooks' study (2001). However, in terms of implications for research, knowing how test takers feel about the mediums of test delivery can possibly help the development and implementation of computer-based tests.

5.4 Recommendations for Future Research

The following areas are recommended for further research.

1. Development of computer-adaptive tests (CAT) which assess complex performance of reading abilities by changes in the testing format is worth explored. Further research may

investigate the influence of both the quality of reading performances and the accuracy of proficiency levels.

2. Investigation on whether the adaptive approach will result in the selection of tasks which adequately reflect an examinee's actual abilities, and whether the number and range of tasks will provide sufficient evidence for an accurate level of comprehension.

3. As a result of the growing interest in performance assessment in educational and language assessment (Feldmann and Fish, 1988), a variety of item types must be included for investigating the mode effect on reading comprehension.

4. Investigating whether a very large pool of tasks adequately represents interesting topics for reading levels, in such a way that the adaptive algorithm may select a number of unique tasks close to the examinee's proficiency.

5. In terms of online language testing, the following issues can be investigated: 1) validation procedures for different types of media use, different types of delivery platforms, and the equivalence of test-taking in different environments, 2) the potential, limits, and optimal uses of online language tests and 3) the possibilities of virtual reality for near-perfect task authenticity and performance-based testing.

6. Investigation into the effects of presentation mode on measuring the reading process. Adequate process measures should be devised and included in empirical studies. Methodologies such as an analysis of eye movement and a verbal protocol analysis and questionnaires may be useful for this purpose.

7. Regarding the comparability studies for second reading tests, the evaluation of the equivalence between computer-based and paper-based tests from various perspectives is needed. Some issues that should be covered in comparability studies may include the possibility of similarity of test constructs, examinees' characteristics and testing conditions.